



# CHRONIC KIDNEY DISEASE

FINDING A  
SUITABLE  
CLASSIFICATION  
MODEL

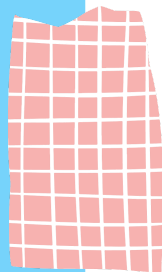
# INTRO

Chronic Kidney Disease is one of the most critical illness nowadays and proper diagnosis is required as soon as possible. Machine learning techniques have become reliable for medical treatment. With the help of a machine learning classifier algorithms, the doctor can detect the disease on time.



Canva

# ABOUT THE ISSUE



---

Chronic kidney Disease (CKD) means your kidneys are damaged and not filtering your blood the way it should. The primary role of kidneys is to filter extra water and waste from your blood to produce urine and if the person has suffered from CKD, it means that wastes are collected in the body. This disease is chronic because of the damage gradually over a long period.

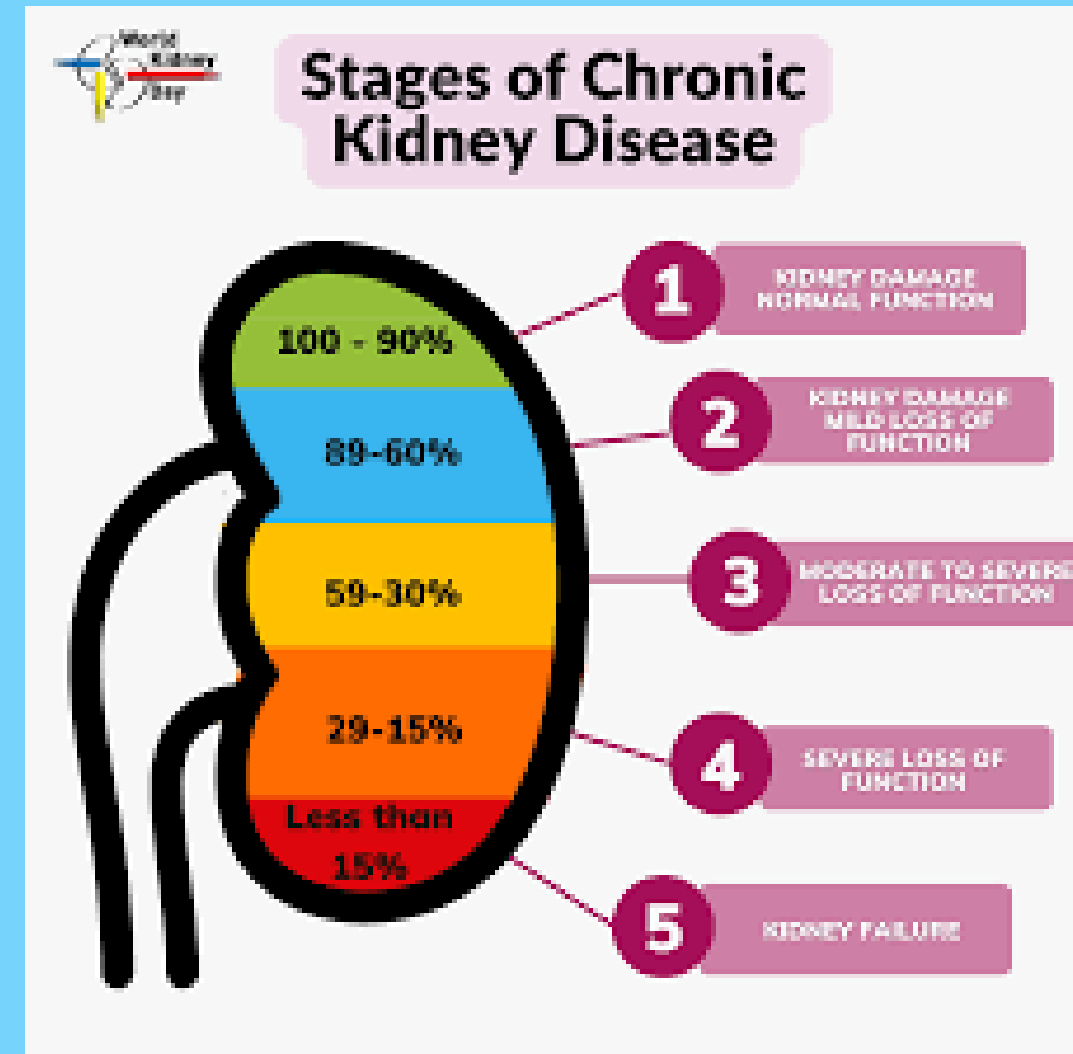
“

**WITH EARLY DETECTION SERIOUS  
ISSUES RELATED TO CKD CAN BE  
AVOIDED. THUS THE ML APPROACH  
CAN HELP US IN EARLY DETECTION  
OF THE DISEASE.**



# AIM

THE AIM OF THIS PROJECT IS TO UNDERSTAND OUR DATA AND USE IT TO CLASSIFY THE PATIENTS INTO TWO CLASSES BASED ON WHETHER THEY HAVE CKD OR NOT. WE HAVE USED 5 ALGORITHMS TO CLASSIFY OUR DATA.



# OUR DATASET

1)THE DATASET CONSISTS OF 25 FEATURES  
WHICH ARE SHOWN IN THE TABLE.

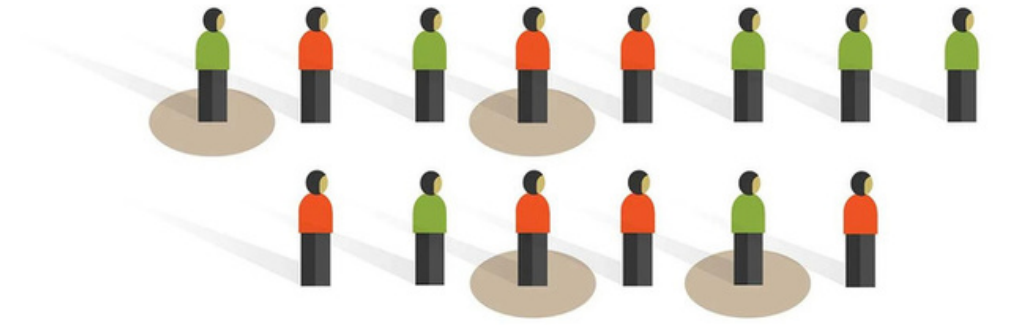
2)Chronic Kidney Disease dataset has been  
taken from the UCI repository.

Sr. No	Attribute Name	Description
1	Age	Patient age (It is in years)
2	Bp	Patient blood pressure (It is in mm/HG)
3	Sg	Patient urine specific gravity
4	Al	Patient albumin ranges from 0-5
5	Su	Patient sugar ranges from 0-5
6	Rbc	Patient red blood cells two value normal and abnormal
7	Pc	Patient pus cell two value normal and abnormal
8	Pcc	Patient pus cell clumps two values present and not present
9	Ba	Patient bacteria two values present and not present
10	Bgr	Patient blood glucose random in mg/dl
11	Bu	Patient blood urea in mg/dl
12	Sc	Patient serum creatinine
13	Sod	Patient sodium
14	Pot	Patient potassium
15	Hemo	Patient hemoglobin (protein molecule in red blood cells)
16	Pcv	Patient packed cell volume % of red blood cells in circulating blood
17	Wc	Patient white blood cell counts in per microliter
18	Rc	Patient red blood cell count in million cells per microliter
19	Htn	Patient hypertension two value Yes and No
20	Dm	Patient diabetes mellitus two value Yes and No
21	Cad	Patient coronary artery disease two value Yes and No
22	Appet	Patient appetite two value good and poor
23	Pe	Patient pedal edema two value Yes and No
24	Ane	Patient anemia two value Yes and No
25	Class	Target Variable (CKD or Not)



# DATA CLEANING

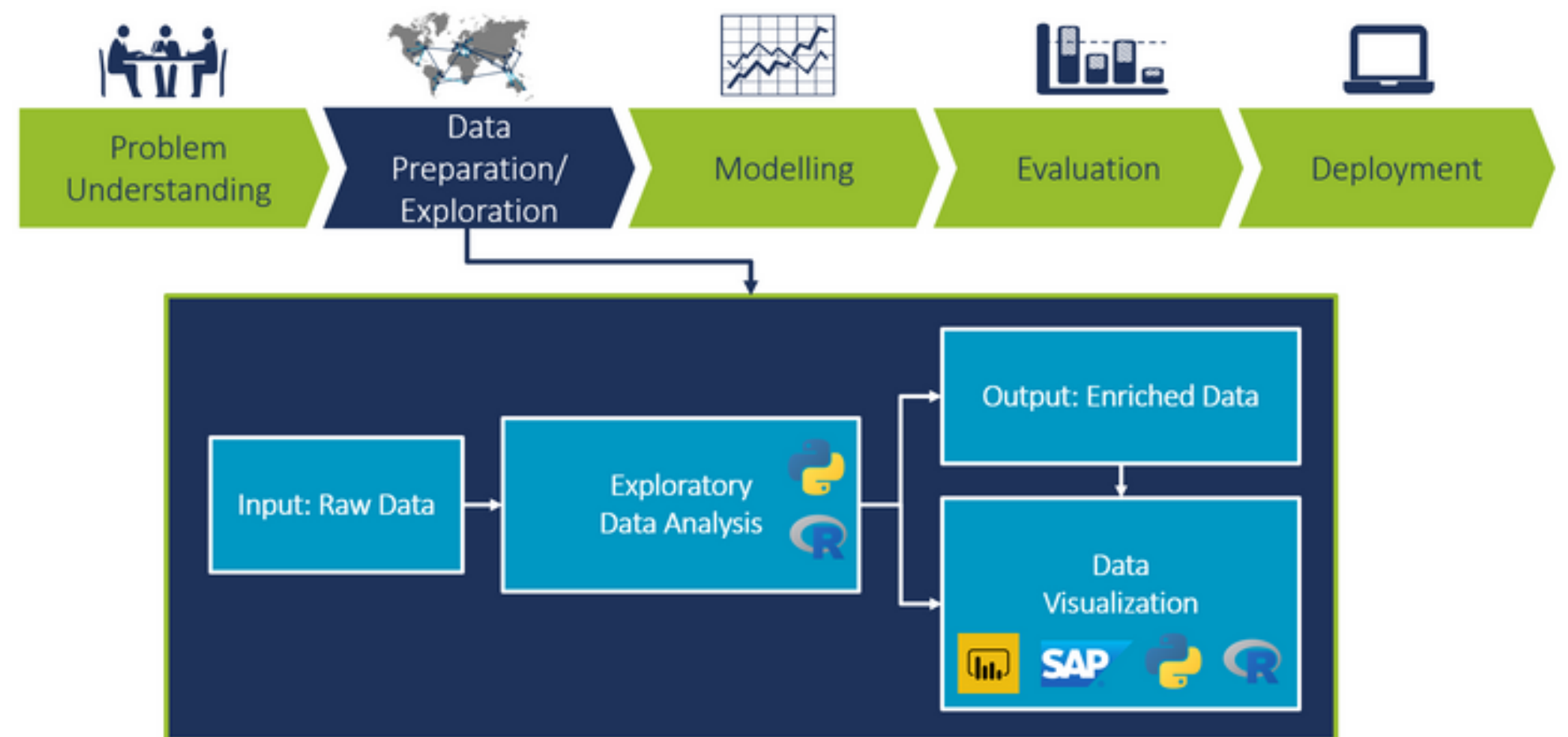
Simple random sampling



- Renaming Columns
- Converting Object data type to numerical data type
- Removing NaN values-Random sampling for higher null values and mean/mode for lower null values

# EXPLORATORY DATA ANALYSIS

- Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.
- Numerical Feature Distribution
- Heat Map
- Violin Graph







# ML MODELLING

1

Feature Encoding

2

Splitting the dataset

3

KNN

4

Decision Tree

5

Logistic Regrassion

6

SVM

7

Random Forest

# FEATURE ENCODING

We use Label encoder as all categorical columns have two categories

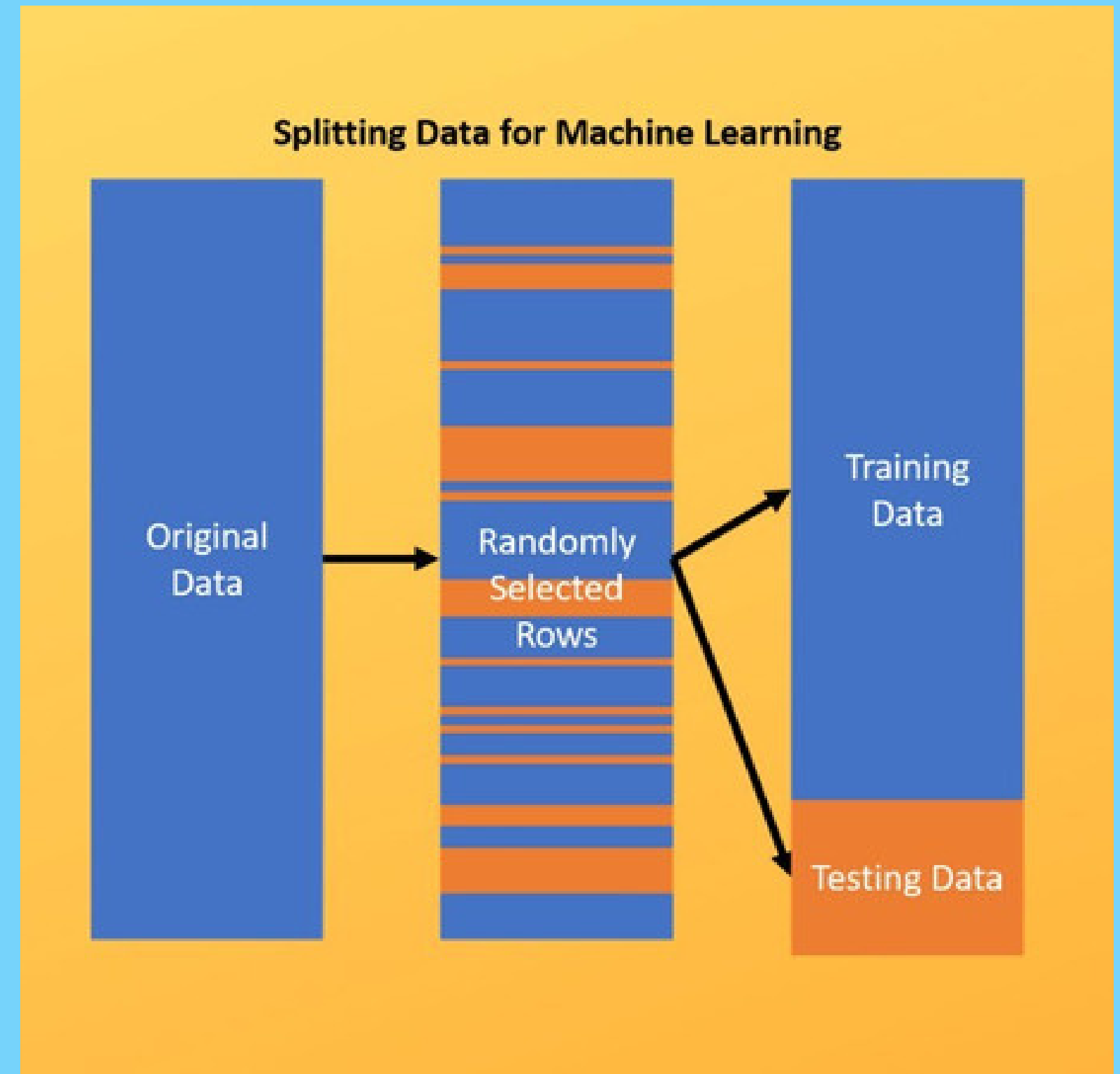
Canva

State (Nominal Scale)	State (Label Encoding)
Maharashtra	3
Tamil Nadu	4
Delhi	0
Karnataka	2
Gujarat	1
Uttar Pradesh	5

# SPLITTING THE DATASET

We Split the dataset into training and testing before using modelling techniques

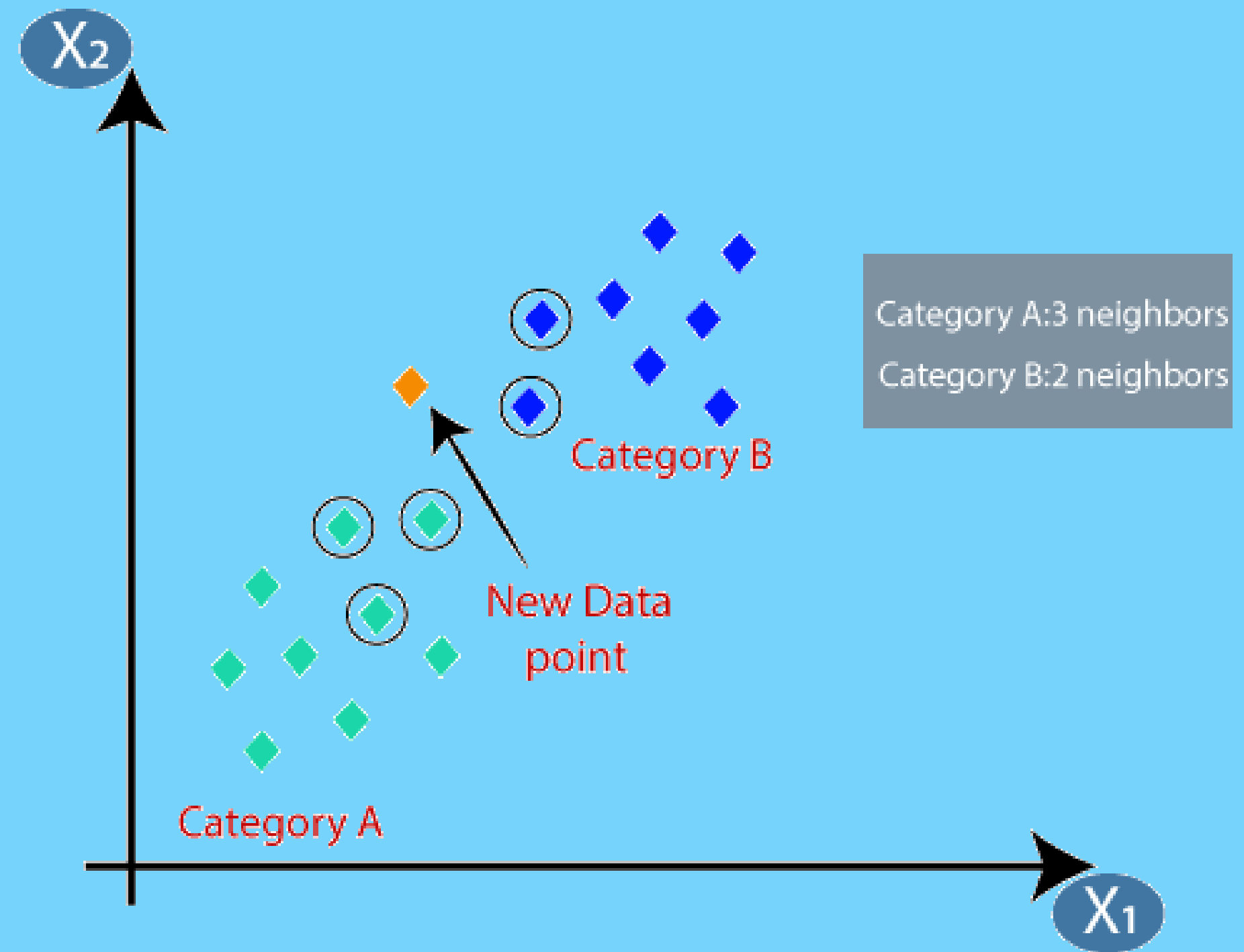
*Canva*



# KNN

- KNN classifies the new data points based on the similarity measure of the earlier stored data points.
- Training Accuracy-78%
- Testing Accuracy-68%

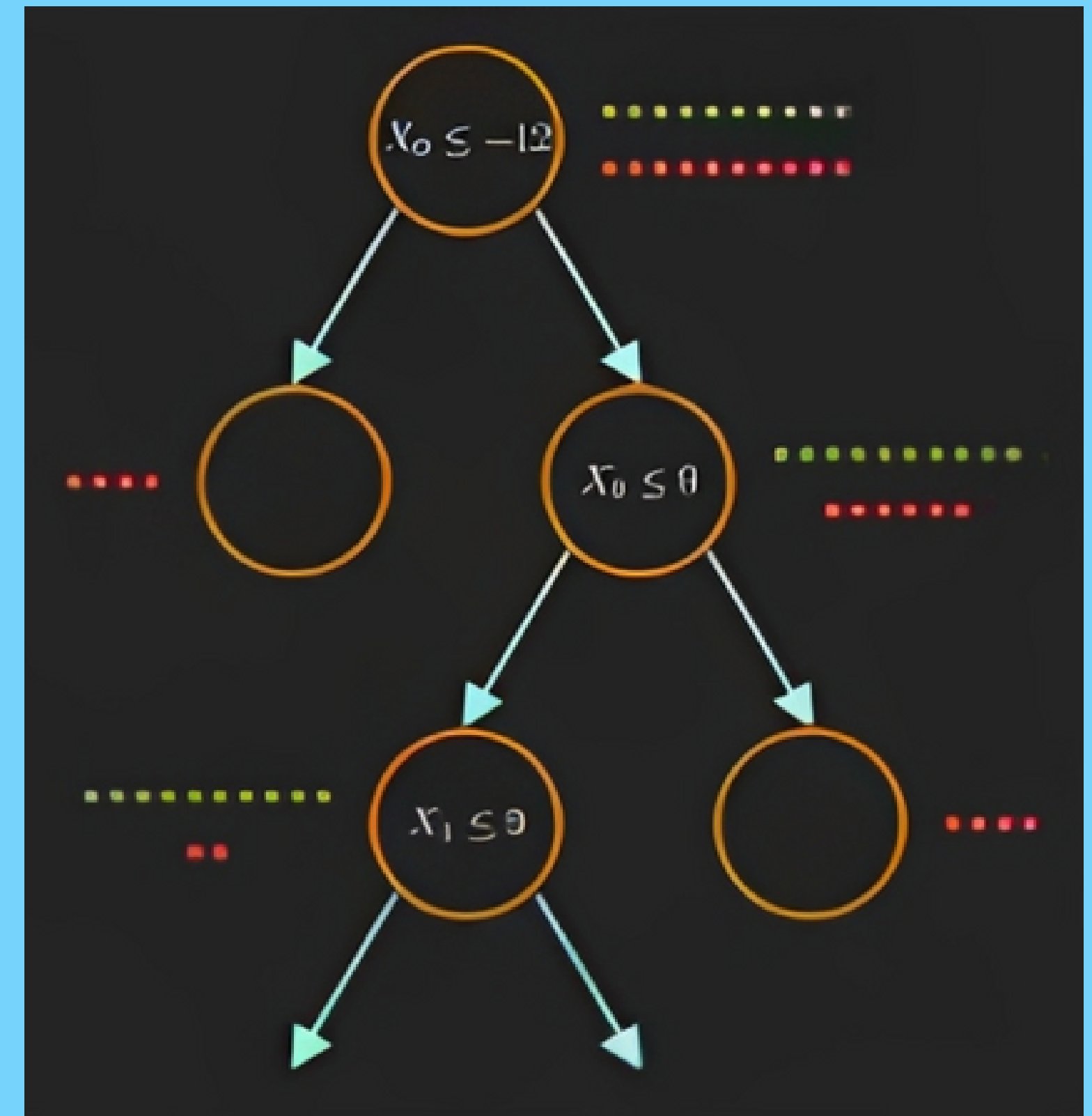
Canva



# DECISION TREE

- Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter.
- Training Accuracy-100%
- Testing Accuracy-96%

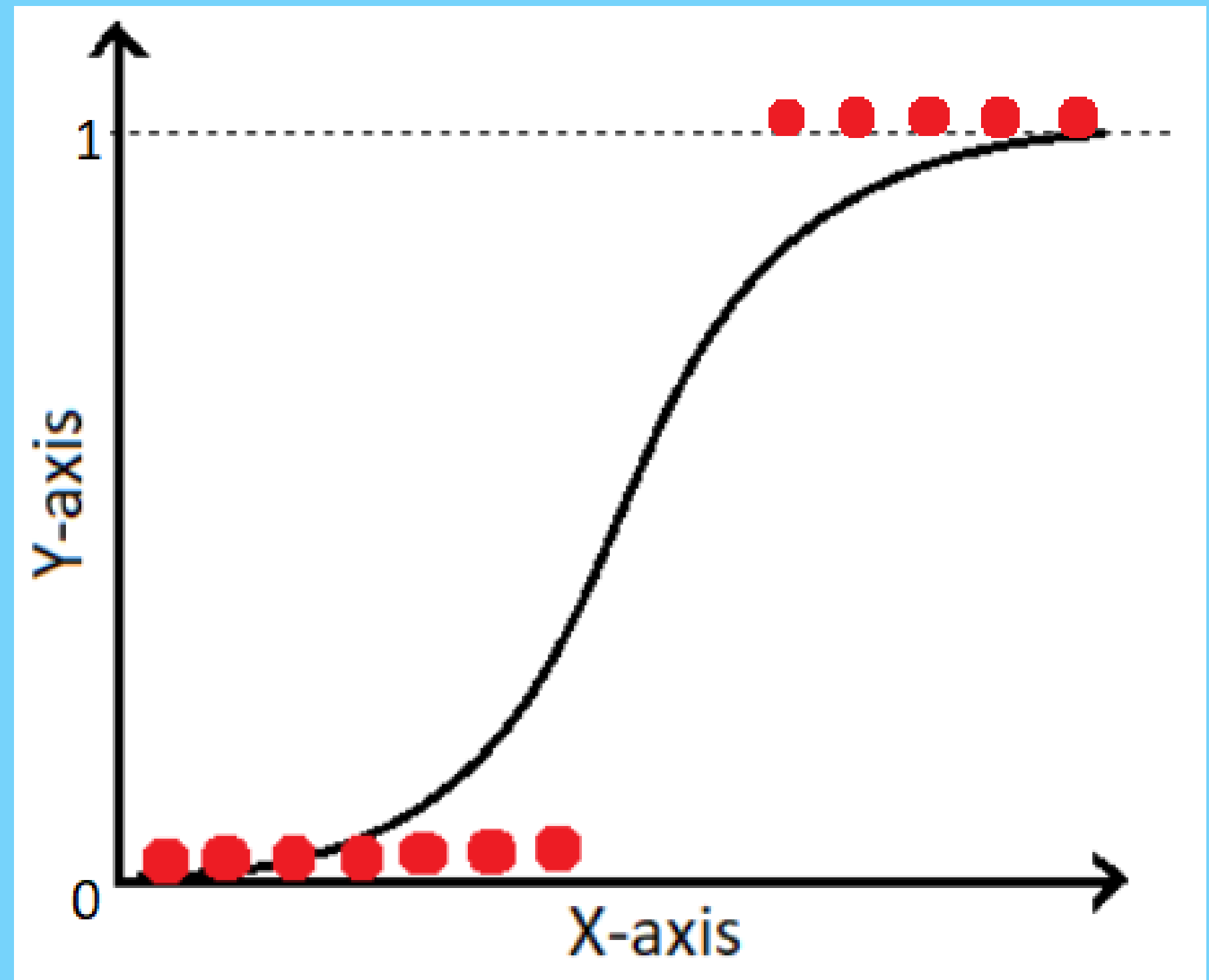
*Carollee*



# LOGISTIC REGRESSION

- Logistic Regression is a statistical model which predicts probability of an event occurring based on dataset of independent variables
- Training Accuracy-90%
- Testing Accuracy-88%

Canva

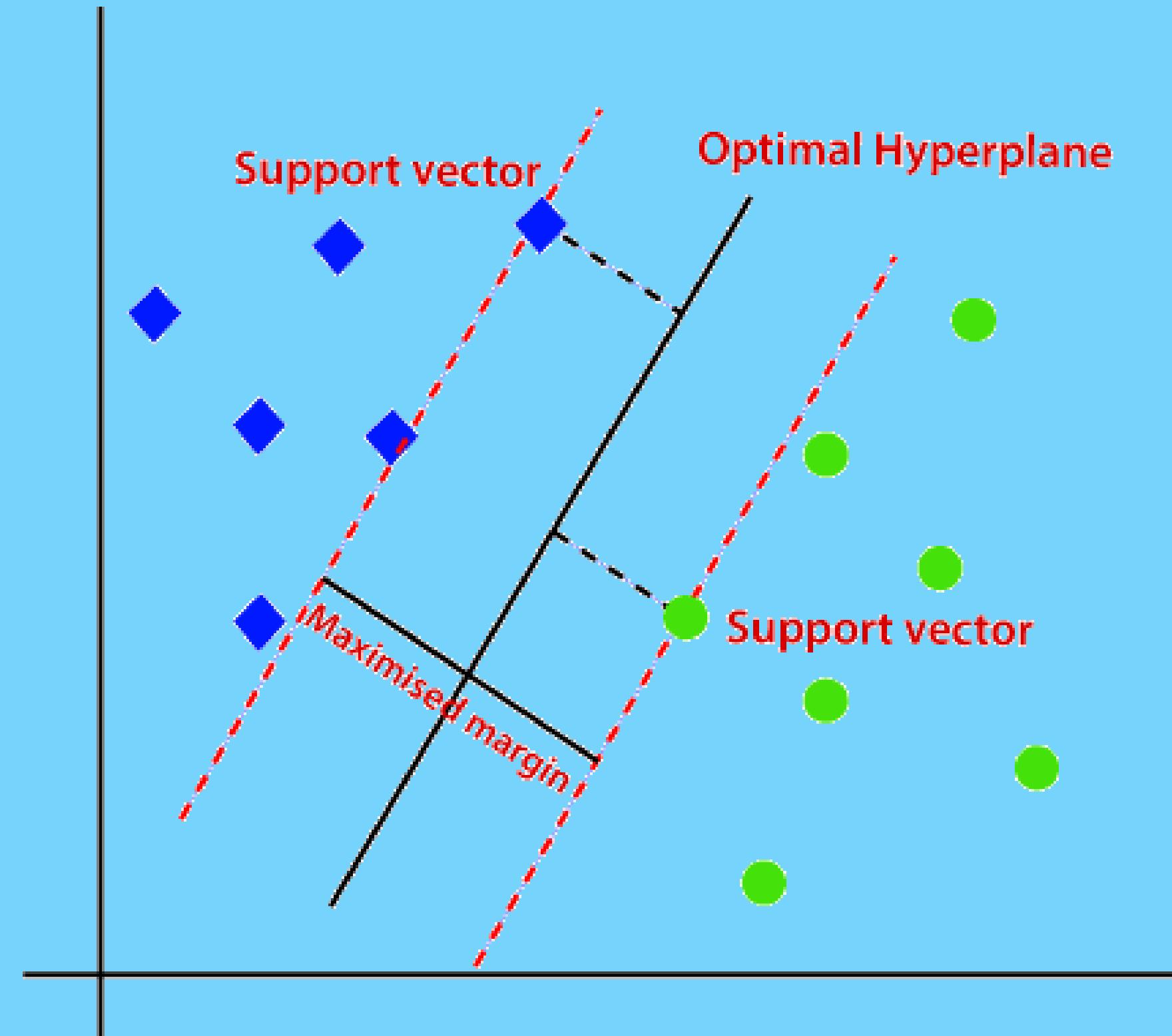




# SVM

- A support vector machine (SVM) is a type of deep learning algorithm that performs supervised learning for classification or regression of data groups.
- Training Accuracy-97%
- Testing Accuracy-94%

Canva

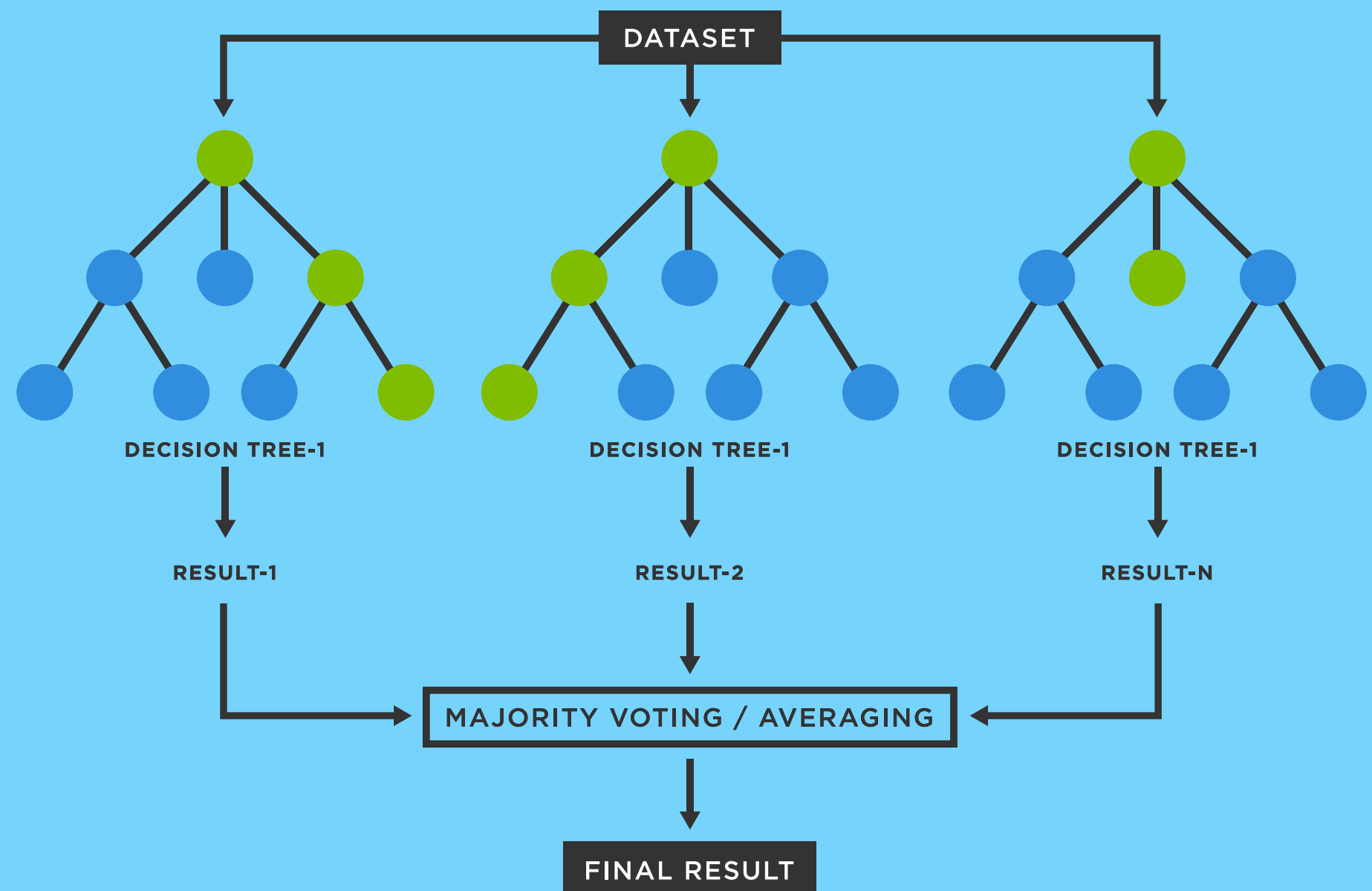


# RANDOM FOREST

100

- The random forest is a classification algorithm consisting of many decisions trees. It uses bagging to try and create uncorrelated forest of trees
- Training Accuracy-100%
- Testing Accuracy-98%

Canva



# WHAT IS SMOTE?

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling.

Canva

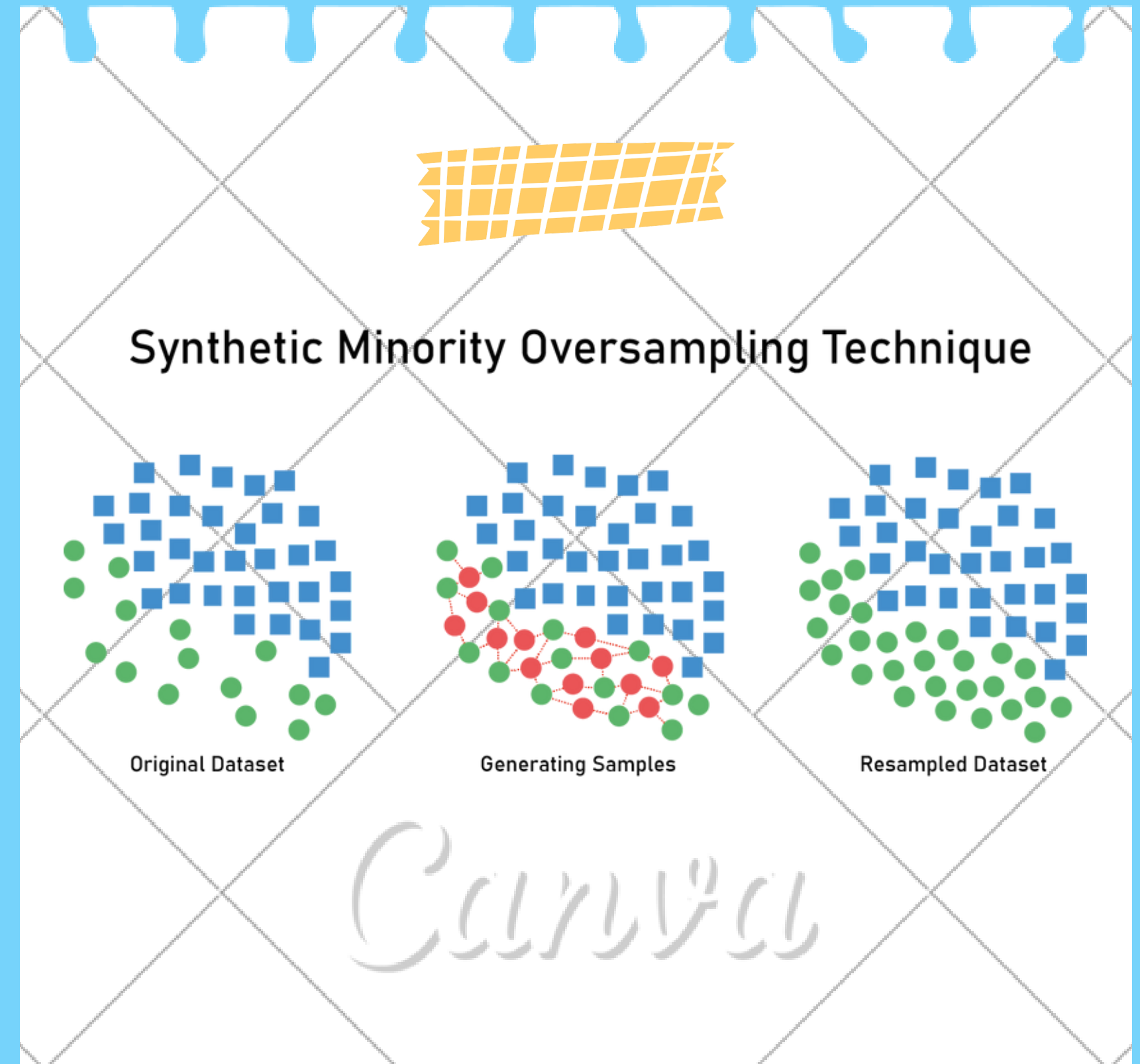
SMOTE works by utilizing a k-nearest neighbour algorithm to create synthetic data. SMOTE first starts by choosing random data from the minority class, then k-nearest neighbors from the data are set. Synthetic data would then be made between the random data and the randomly selected k-nearest neighbor. We have used smote to generate more samples and then tested the algorithms again.

Canva

# RESULTS AFTER SMOTE

After applying smote we observed the following accuracies:

1. KNN- 70%
2. Decision Tree- 96.5%
3. Logistic Regression- 89%
4. SVM- 94%
5. Random Forest- 98%



# **CONCLUSION**

We have achieved a successful model that can classify the patients into whether they have ckd or not. We have observed that we get the highest accuracy for Random Forest followed by decision tree and SVM.

*Canva*

Key Improvements can be:

1. Feature selection based on correlation
2. we used random sampling to fill NaN values instead of just dropping the rows.

*Canva*



**THANK  
YOU!**

**MADE BY:  
ARYAN SINGHAI (20UCS036)  
ARNAV ARORA (20UCS026)**

Have a  
great day  
ahead.